



Key Considerations for Measuring Allelic Expression on a Genomic Scale Using High-throughput Sequencing

Citation

Fontanillas, Pierre, Christian R. Landry, Patricia J. Wittkopp, Carsten Russ, Jonathan D. Gruber, Chad Nusbaum, and Daniel L. Hartl. 2010. Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing. *Next Generation Molecular Ecology*. Special Issue. *Molecular Ecology* 19:212-227.

Published Version

10.1111/j.1365-294X.2010.04472.x

Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:10405235>

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

1 **Key considerations for measuring allelic expression on a genomic scale**
2 **using high-throughput sequencing**

3

4 Pierre Fontanillas^{1,2}, Christian R. Landry^{2,3,4}, Patricia J. Wittkopp⁵, Carsten Russ⁶,
5 Jonathan D. Gruber⁵, Chad Nusbaum⁶, Daniel L. Hartl²

6

7

8 ¹ Department of Ecology and Evolution, University of Lausanne, Le Biophore, CH-1015
9 Lausanne (Switzerland).

10

11 ² Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity
12 Avenue, Cambridge MA 01398 (USA).

13

14 ³ Département de Biochimie, Université de Montréal, C.P. 6128, Succ. Centre-Ville,
15 Montréal (Canada).

16

17 ⁴ Institut de Biologie Intégrative et des Systèmes (IBIS), Département de Biologie,
18 Université Laval, G1V 0A6, Québec (Canada).

19

20 ⁵ Department of Ecology and Evolutionary Biology, Department of Molecular, Cellular,
21 and Developmental Biology, University of Michigan, 830 North University Avenue, Ann
22 Arbor, MI 48109 (USA).

23

1 ⁶ Broad Institute, 7 Cambridge Center, Cambridge, MA 02142 (USA).

2

3 Corresponding author:

4 Pierre Fontanillas

5 Department of Ecology and Evolution, University of Lausanne, Le Biophore, CH-1015

6 Lausanne (Switzerland).

7 pierre.fontanillas@unil.ch

8 Phone: +41 21 692 4204

9 Fax: +41 21 692 4165

10

11 Running title: Allelic expression and high-throughput sequencing.

12

13 Keywords: gene expression, *cis*-regulation, *Drosophila melanogaster*, *Drosophila*
14 *simulans*, hybrids.

15

1 **Abstract:**

2 Differences in gene expression are thought to be an important source of
3 phenotypic diversity, so dissecting the genetic components of natural variation in gene
4 expression is important for understanding the evolutionary mechanisms that lead to
5 adaptation. Gene expression is a complex trait that, in diploid organisms, results from
6 transcription of both maternal and paternal alleles. Directly measuring allelic expression
7 rather than total gene expression offers greater insight into regulatory variation. The
8 recent emergence of high-throughput sequencing offers an unprecedented opportunity to
9 study allelic transcription at a genomic scale for virtually any species. By sequencing
10 transcript pools derived from heterozygous individuals, estimates of allelic expression
11 can be directly obtained. The statistical power of this approach is influenced by the
12 number of transcripts sequenced and the ability to unambiguously assign individual
13 sequence fragments to specific alleles on the basis of transcribed nucleotide
14 polymorphisms. Here, using mathematical modeling and computer simulations, we
15 determine the minimum sequencing depth required to accurately measure relative allelic
16 expression and detect allelic imbalance via high-throughput sequencing under a variety of
17 conditions. We conclude that, within a species, an minimum of 500 to 1000 sequencing
18 reads per gene are needed to test for allelic imbalance, and consequently, at least five to
19 10 millions reads are required for studying a genome expressing 10,000 genes. Finally,
20 using 454 sequencing, we illustrate an application of allelic expression by testing for *cis*-
21 regulatory divergence between closely related *Drosophila* species.

1 A major challenge in evolutionary biology today is understanding the genetic and
2 molecular mechanisms that give rise to phenotypic differences within and between
3 species. Such differences can arise from mutations affecting the function of gene products
4 (i.e., proteins or RNAs) or mutations that affect expression of these genes. Historically,
5 researchers have looked almost exclusively for (and often found) changes in protein
6 coding regions that appeared to contribute to phenotypic evolution; however, during the
7 last decade, there has been a dramatic increase in the number of studies showing that
8 changes affecting gene regulation can also bring about diversity in ecologically relevant
9 traits that affect behavior, physiology, and morphology (e.g. Duda & Remigio 2008;
10 Giger *et al.* 2008; Voelckel *et al.* 2008; see also for reviews Wray 2007; Hoekstra &
11 Coyne 2007; Stern & Orgogozo 2008; Pennisi 2008).

12 Studies of gene expression have become routine with the development of
13 techniques that quantify transcript abundance in a high-throughput way. Microarray
14 studies, in particular, have produced valuable catalogs of differences in transcript levels
15 between individuals (Oleksiak *et al.* 2002; Whitehead & Crawford 2006), between
16 species in diverse taxa (Rifkin *et al.* 2003), and between ecological conditions (Reymond
17 *et al.* 2000; Carsten *et al.* 2005; Derome *et al.* 2006). Such studies also show that inter-
18 individual differences in gene expression are often highly heritable (Wayne *et al.* 2004;
19 Gibson & Weir 2005; Hugues *et al.* 2006; Lemos *et al.* 2008; Ayroles *et al.* 2009).

20 Because of this heritability, quantitative trait locus (QTL) mapping can be
21 combined with microarray analysis to investigate the genetic basis of variable gene
22 expression (Vasegami & Primmer 2005). When a QTL affecting a gene's transcription
23 maps close to the affected gene it can be classified as *cis*-acting, while a QTL that maps

1 further away on the same chromosome, or to another chromosome, can be classified as
2 *trans*-acting (Brem *et al.* 2002). However, strictly speaking, “*cis*” describes mutations
3 that affect expression of only the allele on the same chromosome as the mutation,
4 whereas “*trans*” describes mutations that affect allelic expression on both homologous
5 chromosomes. Examples of *cis*-acting sequences include promoters and enhancers, which
6 are typically located close to the gene that they regulate, while examples of *trans*-acting
7 regulators include genes that encode transcription factors, which may be located
8 anywhere in the genome. Classifications of expression QTLs as *cis*- or *trans*-acting based
9 solely on their proximity to the affected gene are therefore only an approximation -- and
10 one that comes with many caveats (Rockman & Kruglyak 2006).

11 Nevertheless, studies mapping expression QTLs suggest that both *cis*- and *trans*-
12 regulatory mutations contribute to transcriptional variation, with a preponderance of
13 expression QTLs appearing to be *cis*-acting (Wayne *et al.* 2004; Hughes *et al.* 2006;
14 Osada *et al.* 2006; Bergen *et al.* 2007; Genissel *et al.* 2008; Gilad *et al.* 2008; Price *et al.*
15 2008; Lemos *et al.* 2008; but see Morley *et al.* 2004), although his methodology generally
16 has less statistical power to detect *trans*-acting than *cis*-acting variants (Cookson *et al.*
17 2009). In addition, QTL mapping studies of variable gene expression require microarrays
18 suitable for studying the species of interest, molecular markers that cover its complete
19 genome, and resources for genotyping these markers in a segregating population. The
20 lack of any one of these things can be a significant impediment for mapping expression
21 QTLs outside of well established genetic model systems.

22 An alternative strategy for studying regulatory variation uses allelic transcript
23 abundance and the fact that *cis*-regulatory mutations have allele-specific effects on gene

1 expression while *trans*-regulatory mutations affect expression of both alleles in a diploid
2 cell (Cowles et al. 2002; Wittkopp et al. 2004). One or more transcribed differences in
3 nucleotide sequence are used to discriminate between transcripts produced by each allele.
4 Asymmetric expression of two alleles, also known as allelic imbalance (AI), that is
5 observed between alleles present in the same cell (i.e., exposed to the same *trans*-
6 regulatory environment) provides direct evidence of *cis*-regulatory differences.
7 Expression differences observed between individuals homozygous for two different
8 alleles that are not also observed between these same alleles in heterozygotes are
9 attributed to *trans*-regulatory differences (Wittkopp et al. 2004).

10 This allele-specific approach has now been used to decompose variable gene
11 expression into its *cis*- and *trans*-regulatory component parts for flies (e.g., Wittkopp *et al.*
12 *al.* 2008), humans (e.g., Pant *et al.* 2006; Serre *et al.* 2008), plants (e.g., deMeaux *et al.*
13 2005, Guo *et al.* 2008), and yeast (Tirosh *et al.* 2009). With the exception of Tirosh *et al.*
14 (2009), who developed custom microarrays, the methods used to measure allelic
15 expression in these studies are not readily scalable to an entire genome. Furthermore,
16 methods used in these studies, including Tirosh *et al.* (2009), require polymorphic sites
17 that differentiate alleles to be known *a priori*. For these reasons, studying allelic
18 expression genome-wide has been impractical for non-model (as well as most model)
19 species.

20 Next generation sequencing technologies have the potential to revolutionize
21 studies of allelic expression. Because they obviate the need for *a priori* sequence
22 information, molecular markers, and locus-specific genotyping assays, next generation
23 sequencing methods can measure allelic abundance at a genomic level in virtually any

1 species. Only transcribed nucleotide differences between alleles and sufficient
2 sequencing depth for detecting AI are required. For these reasons, we expect
3 measurements of allelic expression based on next generation sequencing will soon be
4 acquired by many researchers, not only to disentangle *cis*- and *trans*-regulatory variation,
5 but also to quantify the heritability of gene expression, examine dominance among
6 regulatory alleles, evaluate their contribution to morphological, physiological, or
7 behavioral changes, and reveal patterns of allelic variation within and between species.

8 Not surprisingly, the benefits of next generation sequencing come with a price –
9 and often a high one. A single “run” of high-throughput sequencing can provide up to
10 hundreds of millions of sequences, but currently costs thousands of dollars. The precise
11 cost per base differs among technologies, as does the length of each sequenced fragment
12 and the total number of sequences collected. Because of this cost, careful experimental
13 design that maximizes the data per dollar for allelic expression studies using next
14 generation sequencing is critical. Optimal experimental design is particularly paramount
15 for studies in molecular ecology that seek to examine allelic expression in multiple
16 individuals, species, or environmental conditions.

17 In this paper, we use mathematical modeling and computer simulations to identify
18 critical parameters affecting measurements of allelic expression and the detection of AI
19 with high-throughput sequencing. We show that the statistical power of this method
20 depends upon four crucial parameters (Fig. 1): sequence divergence between alleles, the
21 relative transcript abundance, the average read length (i.e. amount of transcript
22 sequenced), and sequencing depth (i.e. average number of reads per gene). The latter two
23 parameters determine the number of sequencing reads expected to map to each gene. The

1 former two parameters determine the proportion of sequence reads per gene that are
2 informative for allelic expression (i.e., contain one or more single nucleotide
3 polymorphisms (SNPs) that allow reads to be unambiguously assigned to an allele). We
4 show that this probability is strongly affected by the location of SNPs within an mRNA
5 as well as by the way in which the cDNA library is prepared for sequencing. Here, we
6 derive a mathematical model that determines the minimum number of reads required to
7 test for significant AI given various levels of sequence divergence, read lengths, and
8 distributions of relative transcript abundance, and we compare these results with
9 simulations. Finally, to illustrate the potential of this approach, we describe an empirical
10 study using measurements of allelic expression in F₁ hybrids between *Drosophila*
11 *melanogaster* and *D. simulans* obtained using 454 sequencing (Roche 454 Life Sciences).

1 **Quantifying allelic expression using short-read sequencing: expectations and** 2 **statistical power**

3

4 Despite significant improvements in read length since their release, current high-
5 throughput sequencing technologies (e.g., Illumina Solexa, Roche 454 Life Sciences, ABI
6 Solid) remain “short-read” (i.e., < 500 bp) sequencing methods. The length of sequences
7 is particularly important for studying allelic expression because, to be informative, a
8 sequencing read must include one or more SNPs that discriminate between alleles. If the
9 distance between heterozygous sites within a transcript is greater than the average
10 sequence length, many reads will include only invariant sites and thus be uninformative
11 for allelic expression. A single “run” of next generation sequencing generates hundreds of
12 thousands to millions of sequencing reads, but only the subset of reads that are
13 informative contribute to estimates of allelic expression. The relative frequency of
14 informative and uninformative reads has a major impact on the statistical power for
15 detecting AI in a given experiment.

16 In the following section, we derive the probability of obtaining informative and
17 uninformative sequences depending on the number of total sequencing reads, the read
18 length, and the sequence divergence between alleles. The model assumes that informative
19 reads could be assigned unambiguously to alleles: it implies that allelic reference
20 sequences are known and read mapping procedures are without errors. For very short
21 reads, alignments against reference genome can be sensitive to mismatches but a large
22 number of new algorithms have been recently released to deal with this issue (see
23 Bateman & Quackenbush 2009; Kofler *et al.* 2009). Nevertheless, in our model,
24 confidence of allele assignment can be controlled for by the number of SNP required (see

below). We then relax the model assumptions and discuss possibilities of measuring allelic expression without reference genomes or *a priori* sequences. We show that the total number of sequence reads, in combination with the distribution of transcript abundance among genes, determines the sequencing depth (i.e., number of reads) needed per gene. The read length and sequence divergence define the likelihood of sampling one or more transcribed SNPs, which in turn determines the proportion of informative reads. We explore a wide range of parameter values that should encompass most biological comparisons within and between closely related species. These conditions also reflect the current and anticipated output of multiple high-throughput sequencing technologies. The ultimate goal of this work is to provide guidance for researchers designing allelic expression experiments in their favorite system.

The relationship between sequence divergence and read length

When quantifying total levels of transcript abundance by sequencing cDNA, reads from anywhere within the transcript are informative as long as they are long enough to map unambiguously to a single site in the genome (Torres et al. 2008; Wang *et al.* 2009). By contrast, when measuring allelic expression, only the subset of these reads containing one or more SNPs that distinguish transcripts derived from different alleles of the same gene are informative. Therefore, read length is a critical parameter for sequencing-based studies of allelic expression: longer reads increase the probability of sequencing a polymorphic site. Sequence divergence between the two alleles under study must also be considered: greater divergence increases the probability of sequencing polymorphic sites for a given read length.

1 More formally, the probability of sampling at least x SNPs in a read of length l
 2 can be approximated by:

$$3 \quad \Pr(X \geq x) = \sum_{k=x}^l \binom{l}{k} d^k (1-d)^{l-k}, \quad (1)$$

4 where d is the sequence divergence (i.e. the probability of observing a SNP at each
 5 nucleotide position). The probability of obtaining exactly n informative reads with at least
 6 x SNPs in s randomly sampled reads from a heterozygous gene is given by the probability
 7 mass function:

$$8 \quad \Pr(Y = n) = \binom{s}{n} \Pr(X \geq x)^n (1 - \Pr(X \geq x))^{s-n}. \quad (2)$$

9 The mean and the variance of this distribution are

$$10 \quad E(Y) = s \Pr(X \geq x) \quad (3)$$

11 and,

$$12 \quad \text{Var}(Y) = s \Pr(X \geq x) (1 - \Pr(X \geq x)). \quad (4)$$

13 Consequently, $\Pr(X \geq x)$ is the expected proportion of informative reads in a random
 14 sample.

15 If only one SNP is required to assign alleles unambiguously, Eq. 1 becomes:

$$16 \quad \Pr(X \geq 1) = 1 - (1-d)^l. \quad (5)$$

17 For this special case, the minimum read length required to observe a minimum proportion
 18 of $\Pr(X \geq 1)$ reads with at least one SNP is

$$19 \quad l = \frac{\text{Log}(1 - \Pr(X \geq 1))}{\text{Log}(1-d)}. \quad (6)$$

20 Hence, for mRNA sequences that differ at 0.5% of sites, sequence reads longer than 138
 21 bp are needed in order to obtain $\geq 50\%$ of reads with at least one SNP. However, in some

1 cases, more than one SNP may be required to unequivocally discriminate between the
2 two alleles. For example, depending on the specific experiment, ancestral polymorphism,
3 recombination, homoplasy, or errors in transcription or sequencing could lower the
4 confidence of assignments only based on a single SNP. Therefore, we also calculated the
5 expected proportion of informative reads when more than one SNP is required for allele
6 assignment. By applying Eq. 2, we show that this parameter diminishes rapidly when the
7 minimum number of required SNPs increases (Fig. 2A, B). Sequences with little genetic
8 divergence (i.e., $< 1\%$) are particularly sensitive to the minimum number of SNPs
9 required to discriminate between alleles.

10 To determine how well this mathematical model predicts the proportion of
11 informative reads, we randomly sampled reads *in silico* with lengths ranging from 35 to
12 800 bp from pairs of virtual mRNA sequences 2000 bp long that contained various levels
13 of sequence divergence. SNPs were distributed uniformly across the length of each
14 hypothetical mRNA sequence. In each simulation, we counted the number of informative
15 reads, that is, the number containing one or more polymorphic sites. As shown in Figure
16 2C and 2D, predictions from our model are most accurate for sequence reads that are \leq
17 300 bp. For longer reads, our model underestimates the mean as well as the variance.
18 Large means are inaccurately estimated because Eq. 1 assumes that mRNA molecules
19 have infinite length and consequently, that SNPs can be sampled with replacement. When
20 read lengths converge to mRNA lengths, this assumption then becomes invalid and Eq. 1
21 should be replaced by a hypergeometric form that includes a parameter for mRNA length
22 (see legend for Fig. S1). Although this hypergeometric model is indeed more accurate
23 (compare Fig. 2 to Fig. S1), it is also more complex mathematically. For most

1 experimental design applications, we anticipate that the simpler binomial model will be
 2 sufficient.

3 Note that both the binomial (Fig. 2) and hypergeometric (Fig. S1) models
 4 underestimate the variance observed in the simulation study. This is because these models
 5 do not take into account the location of polymorphic sites within each mRNA. The
 6 probability of sampling a SNP dependent upon its position in the mRNA (Supplementary
 7 Fig. S2) can be described as:

$$8 \quad \Pr(B = i) = \frac{w_i}{(l_g - l + 1)l}, \quad \begin{cases} 1 \leq i < l_r : w_i = i \\ l_g - l_r + 1 < i \leq l_g : w_i = l_g - i \\ l_r \leq i \leq l_g - l_r + 1 : w_i = l_r \end{cases} \quad (7)$$

9 where i is the nucleotide position in the mRNA sequence, l_g the mRNA length, and

$$10 \quad \begin{cases} l \leq \frac{l_g}{2} : l_r = l \\ l > \frac{l_g}{2} : l_r = l_g - l + 1 \end{cases}.$$

11 The mean of this probability distribution is

$$12 \quad E(B) = \frac{1}{l_g}, \quad (8)$$

13 and its variance is

$$14 \quad Var(B) = \left(l_g \frac{2l_r - 1}{3l_r(l_r - 1)} - 1 \right) \left(\frac{l_r}{(l_g - l + 1)l} - \frac{1}{l_g} \right)^2. \quad (9)$$

15 Hence, considering the location of a SNP requires a much more complex model and only
 16 special cases can be easily derived. For instance, for the simple situation of an mRNA
 17 sequence containing one and only one SNP, the probability of sampling informative reads
 18 becomes

$$19 \quad \Pr(X = 1) = \frac{l}{l_g}. \quad (10)$$

The model described above treats the read length l as a fixed value even though (after filtering for base quality) all sequencing technologies produce reads with a range of alignable lengths. To incorporate this element of the data, we represent the distribution of l as $\Pr(l)$, and incorporate this new assumption into Eq. 1, resulting in

$$\Pr(X \geq x) = \sum_{l=1}^{+\infty} \Pr(l) \sum_{k=x}^l \binom{l}{k} d^k (1-d)^{l-k} . \quad (11)$$

Surprisingly, incorporating variable read lengths into the model had little effect on the probability of sampling x SNPs in a read. For instance, assuming a Poisson distribution of read lengths, Eq. 5 becomes

$$\Pr(X \geq 1) = 1 - e^{-dl} . \quad (12)$$

Finally, the proportion of genes with more than n informative reads can be estimated by:

$$\Pr(I \geq n) = \sum_{t=n}^{+\infty} \Pr(T=t) \left(1 - \sum_{j=1}^{n-1} \binom{t}{j} \Pr(X \geq x)^j (1 - \Pr(X \geq x))^{t-j} \right), \quad (13)$$

where $\Pr(T=t)$ is the distribution of transcript levels t across genome. This distribution can be empirically determined or approximated by either discrete decay or power law functions (e.g. Ogasawara *et al.* 2003). Assuming a geometric decay and using Eq. 5, we can rewrite Eq. 13 as:

$$\Pr(I \geq n) = \sum_{t=n} \frac{1}{\bar{T}} \left(1 - \frac{1}{\bar{T}} \right)^t \left(1 - \sum_{j=1}^{n-1} \binom{t}{j} (1 - (1 - \bar{d})^l)^j ((1 - \bar{d})^l)^{t-j} \right), \quad (14)$$

where \bar{T} is the mean number of transcripts per gene (i.e. the sequencing depth) and \bar{d} is the mean sequence divergence. Although Eq. 14 lacks some of the variance in parameter values discussed above (i.e., read length and sequence divergence distributions as well as

1 SNP location within a sequence), the model nevertheless does an excellent job of
2 predicting the simulated proportion of genes with more than n informative reads (Fig. 3).

3 Therefore, for a given number of transcribed genes and total number of
4 sequencing reads (from which \bar{T} is derived), as well as a particular mean sequence
5 divergence and mean read length, the number of genes expected to have more than n
6 informative reads can be robustly predicted. This number (n) is critical for designing
7 allelic expression experiments using next generation sequencing because it directly
8 determines the statistical power for detecting AI.

9

10 *Allelic imbalance and statistical power*

11

12 If we specifically define AI as the ratio of allelic transcript abundance (i.e.
13 number of transcripts from allele 1 divided by the number of transcripts from allele 2),
14 then the probability of observing n_{a1} reads from allele 1 in a sample of n informative
15 reads is given by:

$$16 \quad \Pr(N = n_{a1}) = \binom{n}{n_{a1}} \left(\frac{1}{AI + 1} \right)^{n_{a1}} \left(1 - \frac{1}{AI + 1} \right)^{n - n_{a1}}. \quad (15)$$

17 Using this equation, we can determine whether an observed AI value for a gene is
18 significantly different from a null hypothesis of no difference in allelic expression (AI =
19 1). As shown in Figure 4, statistical power rises quickly when the true value of AI is ≥ 2
20 (e.g., a sample of only 50 informative reads provides 60% statistical power). On the other
21 hand, small imbalances (< 1.25 -fold) require more than 500 informative reads to reach
22 this same power. Although statistical power to detect significant AI can be achieved
23 easily with small samples, large samples are generally required to produce reasonably

1 precise estimates of AI, especially when the true value of AI is very large (Fig. S3). Prior
2 studies comparing allelic expression within and between *Drosophila* species observed a
3 median AI of 1.4 for all genes and samples examined and a median AI of 1.7 for cases
4 classified as having significant AI (Wittkopp *et al.* 2008). This suggests that 200 or more
5 informative reads will generally be required to detect a significant and precise AI values
6 using high-throughput sequencing.

7

8 *Read sampling strategies*

9

10 The number of transcribed genes and the mean sequence divergence are both
11 dictated by the species and genotypes under study and cannot be altered. Furthermore, the
12 researcher can affect the average length of sequence reads only in a limited way by
13 choosing one next generation sequencing technology instead of another. The aspect of a
14 sequencing-based allelic expression experiment that the researcher has the most control
15 over is the preparation of cDNA (or equivalent) libraries used for sequencing. Two
16 general types of cDNA libraries can be used for measuring allelic expression. The first is
17 essentially a shotgun library, in which fragments are randomly sampled from the
18 transcriptome. The second is more targeted, containing fragments only from a
19 predetermined region of each transcript. For instance, the 5' or 3' end can be
20 systematically sampled from each transcript (e.g. Gowda *et al.* 2006). The sampling
21 strategy should be chosen carefully because, as shown in Eq. 7 and Supplementary Fig.
22 S2, the location of SNPs within the cDNA template used for sequencing affects the
23 probability of collecting sequence reads informative for allelic expression.

1 For a given sequence divergence and read length, the sampling strategy does not
2 affect the expected mean proportion of informative reads; however it has a large effect on
3 the variance among genes in the proportion of informative reads (Fig. 5C). With a
4 targeted sampling approach all reads for a given gene will either be informative or
5 uninformative, depending on the position of SNPs in the transcript. Consequently,
6 targeted sampling maximizes the statistical power to detect AI for genes that have at least
7 one polymorphic site in the targeted region, but provides no information about relative
8 allelic expression for genes that lack variation in this region. With random sampling, the
9 situation is reversed: estimates of allelic expression can be obtained for more genes, but
10 the power to detect significant AI for any given gene is reduced. Additionally, under
11 random sampling, gene length affects the relative estimated expression among genes (i.e.,
12 more reads should come from longer genes), but has a negligible effect on measures of
13 relative allelic expression because transcripts from both alleles of a gene are usually the
14 same length. (For this reason, we did not include gene length in the mathematical models
15 described above.)

16 To compare the power of random and targeted sampling, we simulated the
17 proportion of genes with more than 200 informative reads under different conditions. As
18 described above, 200 informative reads per gene provides reasonable statistical power to
19 detect reasonably small difference in allelic expression (see Fig. 4). Results from this
20 simulation show that random sampling generally performs better than targeted sampling,
21 except when sequence divergence and the number of sequences per gene are low, and
22 especially when sequencing reads are short (Fig. 5A, B). As an alternative, a “mixed”
23 sampling scheme that included 50% targeted and 50% randomly located sequence reads

1 produced results most similar to targeted sampling (Fig. 5C). Such a “mixed” sampling
2 strategy could be employed by pair-end sequencing a cDNA library containing fragments
3 of variable length that all share the same 5' or 3' end.

4 Based on these results, we conclude that, for allelic expression experiments with
5 at least moderate sequence depth in species with reference genomes available (that can be
6 used to map sequencing reads to particular genes), a random sampling strategy will
7 almost always provide the most information per unit cost. However, if no reference
8 genome sequence is available, or if sequencing depth is limited, targeted sampling may
9 have advantages that offset the loss of information on a genomic scale. For example,
10 targeted sampling simplifies the process of determining which sequence reads come from
11 the same gene. In the absence of a reference genome, this is especially important because
12 sequence reads must be assembled into (hopefully, gene-specific) contigs *de novo* and all
13 reads from the same gene generated by targeted sampling should overlap. Alternatively if
14 a reference genome is available, but only low depth sequencing is possible, a targeted
15 strategy would provide more accurate estimates of allelic expression for the subset of the
16 genes with SNPs in the targeted regions than random sampling. Finally, although there
17 may be some cases for which “mixed” sampling would be the best choice, in the two
18 scenarios considered above, reducing coverage in the targeted region and distributing
19 some of the reads more evenly across the transcriptome does not offer any clear
20 advantages.

21 22 **Allelic expression in *Drosophila* hybrids**

1 To illustrate one application of allelic expression measurements, we quantified
2 allelic expression in cDNA pools derived from interspecific F₁ hybrids and used these
3 data to test for significant AI (i.e., differences in expression between the maternal and
4 paternal alleles). F₁ hybrids used for this study were produced by crossing *D.*
5 *melanogaster* females and *D. simulans* males, collecting virgin female progeny, and
6 aging them for 5-6 days. After extracting mRNA from these flies, a cDNA pool
7 containing random fragments from the transcriptome was constructed and analyzed using
8 the high-throughput sequencing technology developed by 454 Life Sciences (Roche).

9
10 *Analysis of sequencing statistics, informative reads, and allelic expression*
11

12 In all, 36,855 high-quality 454 sequencing reads were obtained from the random
13 interspecific hybrid cDNA library (Table 1). The average length of these reads was 170
14 bp, with 74 and 250 bp for the 2.5 and 97.5% percentile, respectively. 88% of these reads
15 had at least one homologous sequence in either the *D. melanogaster* or *D. simulans*
16 genome (BLAST, *E-value* < 10⁻⁴), 5% of which showed homology to only one of the two
17 parental genomes. These reads mapped to 5591 genes and 975 intergenic regions.
18 Sequences derived from intergenic regions accounted for 7% of the mapped reads, and
19 might have resulted from transcription of unannotated genes or exons, abnormal splicing
20 that generated unexpected splice junctions, transposable elements, or spurious
21 transcription (e.g. Stolc *et al.* 2004). More than 10 reads matched each of 577 genes,
22 while more than 50 reads matched each of 83 genes. Despite this relatively low
23 sequencing depth, our data appear to correctly, albeit crudely, measure total levels of

1 gene expression: estimates of transcript abundance in F₁ hybrids measured using 454
2 sequencing correlated significantly (Spearman correlation, $\text{Rho} = 0.45$, $P < 0.001$) with
3 expression levels reported for female *D. melanogaster* that were obtained using
4 microarrays (Harbison *et al.* 2005).

5 In order to assign each read to a specific parental allele, we determined the
6 number of variable sites among 454 reads for a particular gene (i.e., SNPs) that were
7 identical to the *D. melanogaster* and *D. simulans* reference genomes. This analysis was
8 performed only for reads that showed a significant alignment with both reference
9 genomes. Sequences were assigned to the species with the highest number of identical
10 SNPs. For instance, if 3 SNPs within a given read were identical to the *D. melanogaster*
11 reference genome and 2 (different) SNPs were identical to the *D. simulans* reference
12 genome, the read was assumed to be derived from the *D. melanogaster* allele. Reads were
13 expected to differ from the reference genomes because the strains of *D. melanogaster* and
14 *D. simulans* used to generate the F₁ hybrids analyzed by 454 sequencing were not the
15 same as the strains sequenced to assemble these reference genomes. As mentioned above,
16 shared recombination, homoplasy, or errors in transcription or sequencing might also
17 complicate the assignment of individual reads to one species or the other. In all, 26,039
18 reads, which is 71% of all high-quality reads, were assigned to one species or the other.
19 At least 10 informative reads were identified for each of 465 genes, with 58 of these
20 genes having more than 50 informative reads each (Fig. 6B). As shown in Figure 4, genes
21 with 50 informative reads had 60% power to detect significant AI of at least 2-fold, while
22 genes with 10 reads had only 20% power to detect changes of the same magnitude.
23 Consistent with both intuitive and mathematical predictions, long reads were more often

1 informative for allele-specific expression than short reads (Fig. 6A), with the average
2 length of uninformative reads only ~70% that of informative reads (Table 1).

3 Intriguingly, more informative reads (53%) were assigned to the *D. melanogaster*
4 allele than to the *D. simulans* allele (13,770 and 12,269 reads, respectively; binomial test:
5 $P < 0.001$). The excess of *D. melanogaster* alleles was greater among sequences from
6 genomic regions annotated as genic than those annotated as intergenic (53.0% vs 50.3%),
7 although this difference was not significant ($\text{Chi}^2 = 3$, $P = 0.08$) and may be an artifact of
8 the much smaller number of reads classified as intergenic (93% vs 7%). The
9 overabundance of *D. melanogaster* alleles was observed not only across the whole
10 transcriptome analysed but also for subsets of genes with different expression levels (Fig.
11 6C). The one (surprising) exception to this was the subset of genes with the highest
12 overall expression levels, which showed an excess of *D. simulans* alleles (Fig. 6C). This
13 shows that the greater abundance of *D. melanogaster* reads observed in F₁ hybrids cannot
14 be explained by higher expression of the *D. melanogaster* allele of only a few highly
15 expressed genes. In addition, this pattern is unlikely to result from poor quality sequences
16 or alignments because (1) BLAST hit lengths are on average identical in both reference
17 genomes, (2) there were on average 4.5 discriminating SNPs per informative read, in both
18 species, which makes nearly all assignments unambiguous, and (3) the trend remained
19 after applying a higher stringency cutoff (≥ 2 SNPs, see Table 1). Based on all of these
20 observations, we conclude that the preferential expression of *D. melanogaster* alleles
21 observed in F₁ hybrids is a real (biological) property of these flies and likely results from
22 a process that affects allelic transcription on a genomic-scale. This expression bias does
23 not appear to be caused by a particular tissue, a particular biological function, or a

particular chromosomal location (data not shown); however, deeper sequence coverage is required to test these hypotheses definitively.

Allelic imbalance in Drosophila hybrids

As described in the introduction, quantifying AI in F₁ hybrids provides a direct readout of relative *cis*-regulatory activity. We tested for significant AI of the 891 genes that show more than 6 informative reads because it is impossible to reach significance ($P < 0.05$) with a binomial test and less than 6 observations. 107 of these 891 genes (12%) showed significant AI based on our data (Supplementary Table S1). For this analysis, we estimate the false discovery rate (i.e. FDR) to be about 17% by resampling our observed distribution of informative read abundance assuming observed AIs are true or equal to 1. On the other hand, we estimated about 13% of false negative AI. Consequently, the analysis suggests that more than 190 genes could show true AI (21% of tested genes). In fact, despite the low level of sequencing depth, we detected significant AI for 35% of genes with more than 50 informative reads and for 14% of genes with more than 10 informative reads. On average, the more abundant allele of a gene was observed 5 times more than the less abundant allele, with a minimum of 1.6 and a maximum of 15 (Supplementary Fig. S4). Finally, and perhaps not surprisingly given the overall excess of reads most similar to *D. melanogaster*, the *D. melanogaster* allele was more abundant than the *D. simulans* allele for 73 (68%) of these 107 genes (binomial test: $P < 0.001$, Fig. 6C).

1 *454 sequencing and pyrosequencing produce similar estimates of allelic expression*

2
3 To determine whether 454 sequencing produces accurate estimates of allelic
4 expression, we used pyrosequencing to independently quantify AI in 14 genes and
5 compared the results to estimates obtained using 454 sequencing. The 14 genes selected
6 for this analysis had allelic expression differences that spanned the range of AI values
7 observed with 454 sequencing (compare Fig. 6D to Fig. S4). Pyrosequencing permits
8 quantifications of allelic expression for individual genes and produces estimates of AI
9 that are consistent with those obtained using both quantitative real-time PCR and
10 microarrays (Wittkopp *et al.* 2006). Pyrosequencing allows high levels of replication for
11 a modest cost, thus very precise estimates of AI can be obtained. Despite the small
12 number of informative 454 sequencing reads for some of the genes analyzed
13 (Supplementary Table S2), we found a significant correlation between estimates of allelic
14 expression obtained using these two techniques (Fig. 6D; Spearman's $\rho = 0.65$, $P <$
15 0.05). The effect of sequencing depth for individual genes on the precision of allelic
16 expression estimates is readily visible by examining the gene-specific binomial sampling
17 confidence intervals shown in Figure 6D. Because precision improves with deeper
18 sequencing coverage, we also examined the correlation between measures of allelic
19 expression for the 8 genes with at least twenty informative reads each. This subset of the
20 data showed a stronger correlation (Spearman's $\rho = 0.74$, $P < 0.05$), as expected.

1

2 **Discussion**

3

4 Less than a decade ago, Gibson (2002) predicted that microarrays would have a
5 tremendous impact on molecular and genetic research at the interface of organismal and
6 population biology. Since then, this methodology has proven to be a formidable tool for
7 examining the extent of gene expression polymorphism within species and divergence
8 between species, as well as for studying the evolutionary processes that generate and act
9 upon regulatory variation. Microarray studies have revealed extensive inter-individual
10 variation in gene expression and showed that this variation is often heritable and its
11 segregation within and between species is primarily affected by neutral drift and
12 stabilizing selection (see for review Whitehead & Crawford 2006). Perhaps surprisingly,
13 very little adaptive regulatory variation has been identified. This may be because it is
14 uncommon or because we do not yet have the correct theoretical framework for
15 distinguishing the effects of neutral and non-neutral evolution (Fay & Wittkopp 2008).

16 High-throughput sequencing technologies are now poised to replace microarrays
17 for measuring gene expression on a genomic scale, especially for evolutionary and
18 ecological studies that require analysis of many species and/or individuals with divergent
19 genotypes. Compared to microarrays, a high-throughput sequencing approach is more
20 complex technologically, yet much simpler statistically and methodologically.
21 Furthermore, it does not require taxon-specific probes or any *a priori* sequence
22 information, making it suitable for measuring expression profiles of virtually any species.
23 Arguably, its greatest advantage relative to microarrays is that both total and allelic

1 measures of gene expression are obtained simultaneously. With measures of allelic
2 expression, there is a closer relationship between an allele's sequence and its activity,
3 making it easier to study the inheritance of gene regulation phenotypes within and
4 between species.

5 Indeed, using allelic expression rather than total gene expression not only
6 facilitates studies exploring the evolutionary consequences of naturally occurring
7 regulatory variation, but also simplifies mapping of regulatory mutations variation, which
8 is necessary to elucidate the genetic architecture of gene expression. Such studies allow
9 researchers to more efficiently identify loci, genes, and molecular processes that play
10 important roles in adaptation. We fully anticipate that studies of allelic expression will
11 soon produce novel insights for the field of molecular ecology, as well as many other
12 fields. We hope that the theoretical and empirical information provided by this study
13 helps researchers design experiments that test specific hypotheses efficiently and cost-
14 effectively.

15 To this end, we have developed a mathematical model that allows researchers to
16 estimate the minimum sequencing depth needed to detect significant AI, based on the
17 average sequence read length of their chosen next generation sequencing technology, the
18 average genetic divergence between alleles under study, the approximate number of
19 transcribed genes, and the anticipated distribution of transcript levels across the genome.
20 For instance, Eq. 14 shows that, for an experiment that uses reads averaging 150 bp and
21 compares allelic expression in a polymorphic species with 0.1% sequence divergence,
22 more than 4,000 sequence reads per gene are needed to achieve 60% statistical power for
23 detecting significantly AI larger than 1.5-fold (see Fig. 3 and 4). For a genome containing

1 10,000 transcribed genes (and assuming an exponential decay distribution for expression
2 levels), such a study would require approximately 40 million reads (Fig. 3). If the average
3 sequence length were increased to 500 bp, the number of sequence reads required would
4 be decreased 4-fold to approximately 10 million reads. With only three million reads, a
5 statistical power of ~20% would be achieved.

6

7 *Caveats and considerations*

8

9 We stress that these predictions should be treated as approximate guidelines only
10 because a number of features of real transcriptomes violate our model and cause our
11 equations to over- or underestimate the requisite number of sequence reads. In the
12 following paragraphs, we discuss four such features and their impact on study design.

13 We modeled the distribution of transcript levels using a standard decay function,
14 yet the distribution of transcript abundances in real organisms will virtually never fit such
15 as function perfectly. In particular, strong deviations in distribution tails, such as many
16 more genes with extremely high or low expression levels, will cause the model to
17 underestimate the required minimum sequencing depth. If the precise distribution of
18 transcript levels is known *a priori* for a particular organism, however, this information
19 can easily be incorporated into Eq. 13 to improve the accuracy of the predictions.

20 Another potential caveat comes from the imprecise construction of cDNA pools.
21 We found that the way sequences are sampled from mRNA transcripts affects the
22 statistical power for a given set of parameters (Fig. 5). We modeled cases only in which
23 cDNA pools contain completely random fragments and perfectly targeted fragments;

1 however, one or more of the technical steps used to create cDNA libraries for sequencing
2 may often introduce imperfections. For example, an unexpected bias may arise during the
3 production of “random” cDNA libraries caused by factors such as the preferential ligation
4 of adapters used for sequencing. Alternatively, targeted libraries may include unwanted
5 sequence fragments that result from reads in untargeted regions. However, our analysis of
6 the “mixed” sampling strategy suggests that small errors in the ultimate composition of
7 cDNA libraries will have little effect on the power to detect AI for most genes. Paired-
8 end sequencing, which is an option now available for some technologies, has no clear
9 advantages for AI analyses, although it could be used to evaluate read and sample quality.

10 Yet another factor to consider carefully is which genes (with respect to expression
11 level) are of utmost interest. For example, if highly transcribed genes are of primary
12 concern, shallow sequencing may be sufficient; increasing sequencing depth has little
13 impact on the statistical power to detect significant AI in these genes. By contrast, if
14 lowly transcribed genes are of most interest, sequencing the entire transcriptome may
15 require such high coverage as to make the experiment impractical. Removing transcripts
16 from genes with the highest expression by subtraction or normalization is one way to
17 increase coverage of the more lowly expressed genes, however, care must be taken not to
18 alter the relative abundance of alleles for the genes of interest. Methods such as
19 quantitative PCR or pyrosequencing may be much more cost-effective and reliable for
20 measuring allelic expression of a few lowly expressed genes.

21 Finally, computer simulations showed that the position of SNPs within a
22 transcript increases variance in the expected number of informative reads per gene. The
23 model assumes however that SNPs are distributed randomly across the length of each

1 mRNA. In reality, polymorphic sites are rarely distributed randomly within a transcript.
2 For instance, they are known to be more frequent in 5' or 3' UTRs than non-synonymous
3 sites (Andolfatto 2005). Differences also exist in the amount of sequence polymorphism
4 and divergence among genes. In general, regions of transcripts with greater sequence
5 variation are more likely to produce informative reads than regions with fewer
6 polymorphic or divergent sites. Finally, our model assumes that transcript abundance and
7 sequence divergence are independent, but in fact, they are often correlated (Subramanian
8 & Kumar 2004).

9

10 *Properties of allelic expression in Drosophila hybrids*

11

12 Interspecific hybrids between *D. melanogaster* and *D. simulans* show extensive
13 misexpression relative to either parental species (Ranz *et al.* 2004). The primary cause of
14 this misexpression is unknown, but it is generally attributed to incompatibilities between
15 genomes that disrupt gene regulatory networks (Landry *et al.* 2007). Consistent with this
16 idea, prior work suggests that the co-evolution of *cis*- and *trans*-regulatory factors may
17 contribute to misexpression when divergent alleles meet in interspecific hybrids (Landry
18 *et al.* 2005).

19 To compare the activity of homologous *cis*-regulatory sequences between species,
20 we used estimates of allelic expression in interspecific F₁ hybrids obtained using 454
21 sequencing. We measured allelic expression for over 5500 of the most highly expressed
22 genes and we were able to test AI in ~900 of these genes. Overall, we detected significant
23 AI for 12% of these genes; among genes with the most informative reads (i.e., ≥ 50 reads

per gene), 35% of genes showed significant AI. Our study almost certainly underestimates the percentage of genes with differences in expression between species-specific alleles in F1 hybrids because even 50 informative reads per gene provides only 25% and 60% statistical power to detect significant AI for genes with true allelic expression differences of 1.5- and 2-fold, respectively. Consistent with this interpretation, prior studies of relative *cis*-regulatory activity between *D. melanogaster* and *D. simulans* reported much higher percentages of genes with significant AI (Lemos *et al.* 2008; Wittkopp *et al.* 2008).

Genomic imbalance in interspecific hybrids

Wide-spread misexpression of genes is common in interspecific hybrids, and the genomic imbalance that we observed in F₁ hybrids between *D. melanogaster* and *D. simulans* alleles may provide insight into the molecular mechanisms responsible for this misexpression. *D. melanogaster* alleles are repeatedly overexpressed in F₁ hybrids despite the fact that both alleles are expected to be regulated by the same pool of maternal and paternal *trans*-acting factors. Long-range chromatin effects that favor expression of *D. melanogaster* chromosomes could contribute to this genomic imbalance, as could interactions between the nuclear genome and cytoplasmic components (especially mitochondria), which were inherited from *D. melanogaster* in the hybrids examined. It will be interesting to see whether these same expression differences exist between the two alleles if they are compared in cells containing only *trans*-acting factors from *D. melanogaster* or *D. simulans*. Such a comparison would provide a direct test for complex

1 interactions among divergent *trans*-acting factors and *cis*-regulatory sequences. Although
2 the prevalence of such interactions remains an open question, a small scale study of *D.*
3 *melanogaster* found no evidence of such *cis*-by-*trans* interactions (Wittkopp *et al.* 2008).

5 *Concluding remarks*

6
7 Comparing the activity of orthologous *cis*-regulatory alleles, as described here for
8 interspecific *Drosophila* hybrids, is but one application of sequence-based measures of
9 allelic expression. Quantifying heritability of gene expression, examining dominance
10 among regulatory alleles, or revealing patterns of allelic variation within and between
11 species are some of additional applications for quantitative measures of relative allelic
12 expression. By identifying parameters critical for such experiments and exploring their
13 impact on possible outputs, we anticipate that results from this study will help researchers
14 optimize their available resources and generate datasets that are best suited for addressing
15 their primary question(s) of interest.

17 **Materials and Methods**

19 *Fly strains, rearing, and crosses*

20
21 F₁ hybrids were produced by crossing four-day-old virgin *D. melanogaster*
22 Canton S females with *D. simulans* C167.4 males. Each mating vial contained 10 females
23 and 10 males. Flies were reared on standard cornmeal medium at 20°C under a 10/14-h

1 light/dark cycle. Hybrid females were collected at emergence and were stored for 5–6
2 days at room temperature and then snap frozen in liquid nitrogen.

3 4 *Preparation of cDNA libraries and 454 sequencing*

5
6 Total RNA was extracted by homogenizing ~500 hybrid females in 4 ml of
7 TRIZOL Reagent (Invitrogen, Carlsbad, CA). We isolated mRNA using an Oligotex
8 Direct mRNA Mini Kit (Qiagen, Valencia, CA). One milligram of mRNA was used for
9 the first-strand cDNA synthesis (Superscript II, Invitrogen, Carlsbad, CA). Reverse
10 transcriptase reactions were performed with biotinylated polyT primer. Second-strand
11 synthesis reactions contained 20 units of DNA ligase, 5 units of RNase H, and 30 units of
12 DNA polymerase I (New England Biolabs, Ipswich, MA). The second-strand reactions
13 were randomly sheared by sonication. The cDNA fragment were blunt-ended with T4
14 polymerase (New England Biolabs, Ipswich, MA). Finally, 3' end cDNA fragments
15 containing the biotinylated polyT primers were removed from cDNA fragment pool by
16 using Dynabeads (Invitrogen, Carlsbad, CA). Sequencing was performed on a GS FLX
17 Instrument following standard protocols (454 Life Science Roche Diagnostics). Sequence
18 reads containing less than 30 bases with high quality score (> 20) were discarded.

19 20 *Data handling and analysis of 454 sequences*

21
22 A custom Perl script was used (a) to BLAST 454 reads against genic and
23 intergenic sequences from *D. melanogaster* (Flybase release 4.1, <http://flybase.org/>) and

1 against the complete genome of *D. simulans* (Apr. 2005 assembly, UCSC Genome
2 Bioinformatics, <http://genome.ucsc.edu/>), (b) to assign to each 454 read to the most
3 probable species (i.e., *D. melanogaster* or *D. simulans*) as well as to a specific gene or
4 intergenic region, and (c) to extract descriptive information from each read: number of
5 basepairs (i.e., sequence length), extent of homology identified by BLAST, genomic
6 position in the *D. melanogaster* and *D. simulans* genomes, number of gaps in the
7 sequence alignments, number of shared SNPs, and number of species-specific SNPs. The
8 first five of these six parameters describe the quality of 454 reads, while the final
9 parameter was used to classify each sequence read as derived from *D. melanogaster* or *D.*
10 *simulans* or to classify it as having an undetermined origin, which was most common for
11 sequence reads that matched either non-polymorphic or extremely polymorphic regions.
12 All statistical analyses and simulations described in this work were performed using R (R
13 Development Core Team, 2005).

14 15 *Pyrosequencing data collection and analysis*

16
17 To validate measures of allelic expression based on 454 sequencing,
18 pyrosequencing assays were performed for 14 genes (Supplementary Table S2) using
19 protocols described in Wittkopp *et al.* (2008). For each gene analyzed, a custom
20 pyrosequencing assay consisting of three unique primers was developed and tested for
21 specificity (primer sequences available upon request). Using these assays, we quantified
22 AI in each of 8 replicate cDNA samples, each independently synthesized from new
23 mRNA extracted from flies collected at the same time as those used for 454 sequencing.

1 Genomic DNA was also extracted from flies obtained in the initial collection and
2 analyzed in quadruplicate using pyrosequencing. From each pyrosequencing reaction, the
3 ratio of *D. melanogaster* to *D. simulans* alleles was calculated as described in Wittkopp *et*
4 *al.* (2008). For each gene, the ratio of relative allelic expression (i.e., *D. melanogaster* /
5 *D. simulans*) was log (base 2) transformed, and the average value from replicate genomic
6 DNA samples was subtracted from the average value of replicate cDNA samples,
7 effectively correcting for any bias in PCR-amplification between alleles (Wittkopp *et al.*
8 2004).
9

1 **References:**

- 2 Andolfatto P (2005) Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* **437**, 1149-
3 1152.
- 4 Ayroles JF, Carbone MA, Stone EA, *et al.* (2009) Systems genetics of complex traits in
5 *Drosophila melanogaster*. *Nature Genetics* **41**, 299-307.
- 6 Bateman A, Quackenbush J (2009) Bioinformatics for Next Generation Sequencing.
7 *Bioinformatics* **25**, 429-429.
- 8 Bergen AW, Baccarelli A, McDaniel TK, *et al.* (2007) *cis* sequence effects on gene expression.
9 *Bmc Genomics* **8**, 296.
- 10 Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional
11 regulation in budding yeast. *Science* **296**, 752-755.
- 12 Carsten LD, Watts T, Markow TA (2005) Gene expression patterns accompanying a dietary shift
13 in *Drosophila melanogaster*. *Molecular Ecology* **14**, 3203-3208.
- 14 Clopper CJ, Pearson ES (1934) The use of confidence or fiducial limits illustrated in the case of
15 the binomial. *Biometrika* **26**, 404-413.
- 16 Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits
17 with global gene expression. *Nature Reviews Genetics* **10**, 184-194.
- 18 Cowles CR, Hirschhorn JN, Altshuler D, Lander ES (2002) Detection of regulatory variation in
19 mouse genes. *Nature Genetics* **32**, 432-437.
- 20 de Meaux J, Goebel U, Pop A, Mitchell-Olds T (2005) Allele-specific assay reveals functional
21 variation in the chalcone synthase promoter of *Arabidopsis thaliana* that is compatible
22 with neutral evolution. *Plant Cell* **17**, 676-690.
- 23 Derome N, Duchesne P, Bernatchez L (2006) Parallelism in gene transcription among sympatric

lake whitefish (*Coregonus clupeaformis* Mitchill) ecotypes. *Molecular Ecology* **15**, 1239-1249.

Duda TF, Remigio EA (2008) Variation and evolution of toxin gene expression patterns of six closely related venomous marine snails. *Molecular Ecology* **17**, 3018-3032.

Fay JC, Wittkopp PJ (2008) Evaluating the role of natural selection in the evolution of gene regulation. *Heredity* **100**, 191-199.

Genissel A, McIntyre LM, Wayne ML, Nuzhdin SV (2008) Cis and trans regulatory effects contribute to natural variation in transcriptome of *Drosophila melanogaster*. *Molecular Biology and Evolution* **25**, 101-110.

Gibson G (2002) Microarrays in ecology and evolution: a preview. *Molecular Ecology* **11**, 17-24.

Gibson G, Weir B (2005) The quantitative genetics of transcription. *Trends in Genetics* **21**, 616-623.

Giger T, Excoffier L, Amstutz U, *et al.* (2008) Population transcriptomics of life-history variation in the genus *Salmo*. *Molecular Ecology* **17**, 3095-3108.

Gilad Y, Rifkin SA, Pritchard JK (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics* **24**, 408-415.

Gowda M, Li HM, Alessi J, *et al.* (2006) Robust analysis of 5'-transcript ends (5'-RATE): a novel technique for transcriptome analysis and genome annotation. *Nucleic Acids Research* **34**, e126.

Guo M, Yang S, Rupe M, *et al.* (2008) Genome-wide allele-specific expression analysis using Massively Parallel Signature Sequencing (MPSS (TM)) Reveals cis- and trans-effects on gene expression in maize hybrid meristem tissue. *Plant Molecular Biology* **66**, 551-563.

1 Harbison ST, Chang S, Kamdar KP, Mackay TFC (2005) Quantitative genomics of starvation
2 stress resistance in *Drosophila*. *Genome Biology* **6**, R36.

3 Hoekstra HE, Coyne JA (2007) The locus of evolution: Evo devo and the genetics of adaptation.
4 *Evolution* **61**, 995-1016.

5 Hughes KA, Ayroles JF, Reedy MM, *et al.* (2006) Segregating variation in the transcriptome:
6 *Cis* regulation and additivity of effects. *Genetics* **173**, 1347-1364.

7 Kofler R, Torres TT, Lelley T, Schlotterer C (2009) PanGEA: Identification of allele specific
8 gene expression using the 454 technology. *Bmc Bioinformatics* **10**, 143.

9 Landry CR, Hartl DL, Ranz JM (2007) Genome clashes in hybrids: insights from gene
10 expression. *Heredity* **99**, 483-493.

11 Landry CR, Wittkopp PJ, Taubes CH, *et al.* (2005) Compensatory cis-trans evolution and the
12 dysregulation of gene expression in interspecific hybrids of *Drosophila*. *Genetics* **171**,
13 1813-1822.

14 Lemos B, Araripe LO, Fontanillas P, Hartl DL (2008) Dominance and the evolutionary
15 accumulation of cis- and trans-effects on gene expression. *Proceedings of the National*
16 *Academy of Sciences of the United States of America* **105**, 14471-14476.

17 Morley M, Molony CM, Weber TM, *et al.* (2004) Genetic analysis of genome-wide variation in
18 human gene expression. *Nature* **430**, 743-747.

19 Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and
20 among natural populations. *Nature Genetics* **32**, 261-266.

21 Osada N, Kohn MH, Wu CI (2006) Genomic inferences of the cis-regulatory nucleotide
22 polymorphisms underlying gene expression differences between *Drosophila*
23 *melanogaster* mating races. *Molecular Biology and Evolution* **23**, 1585-1591.

1 Pant PV, Tao H, Beilharz EJ, *et al.* (2006) Analysis of allelic differential expression in human
2 white blood cells. *Genome Res* **16**, 331-339.

3 Pennisi E (2008) Deciphering the genetics of evolution. *Science* **321**, 760-763.

4 Price AL, Patterson N, Hancks DC, *et al.* (2008) Effects of *cis* and *trans* Genetic Ancestry on
5 Gene Expression in African Americans. *Plos Genetics* **4**, e1000294.

6 Ranz JM, Namgyal K, Gibson G, Hartl DL (2004) Anomalies in the expression profile of
7 interspecific hybrids of *Drosophila melanogaster* and *Drosophila simulans*. *Genome*
8 *Research* **14**, 373-379.

9 Reymond P, Weber H, Damond M, Farmer EE (2000) Differential gene expression in response
10 to mechanical wounding and insect feeding in *Arabidopsis*. *Plant Cell* **12**, 707-719.

11 Rifkin SA, Kim J, White KP (2003) Evolution of gene expression in the *Drosophila*
12 *melanogaster* subgroup. *Nature Genetics* **33**, 138-144.

13 Serre D, Gurd S, Ge B, *et al.* (2008) Differential allelic expression in the human genome: A
14 robust approach to identify genetic and epigenetic *Cis*-acting mechanisms regulating gene
15 expression. *Plos Genetics* **4**, e1000006.

16 Stern DL, Orgogozo V (2008) The loci of evolution: How predictable is genetic evolution ?
17 *Evolution* **62**, 2155-2177.

18 Stolc V, Gauhar Z, Mason C, *et al.* (2004) A gene expression map for the euchromatic genome
19 of *Drosophila melanogaster*. *Science* **306**, 655-660.

20 Subramanian S, Kumar S (2004) Gene expression intensity shapes evolutionary rates of the
21 proteins encoded by the vertebrate genome. *Genetics* **168**, 373-381.

22 Tirosh I, Reikhav S, Levy AA, Barkai N (2009) A Yeast Hybrid Provides Insight into the
23 Evolution of Gene Expression Regulation. *Science* **324**, 659-662.

1 Torres TT, Metta M, Ottenwalder B, Schlotterer C (2008) Gene expression profiling by
2 massively parallel sequencing. *Genome Research* **18**, 172-177.

3 Vasemagi A, Primmer CR (2005) Challenges for identifying functionally important genetic
4 variation: the promise of combining complementary research strategies. *Molecular*
5 *Ecology* **14**, 3623-3642.

6 Voelckel C, Heenan PB, Janssen B, *et al.* (2008) Transcriptional and biochemical signatures of
7 divergence in natural populations of two species of New Zealand alpine Pachycladon.
8 *Molecular Ecology* **17**, 4740-4753.

9 Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics.
10 *Nature Reviews Genetics* **10**, 57-63.

11 Wayne ML, Pan YJ, Nuzhdin SV, McIntyre LM (2004) Additivity and trans-acting effects on
12 gene expression in male *Drosophila simulans*. *Genetics* **168**, 1413-1420.

13 Whitehead A, Crawford DL (2006) Variation within and among species in gene expression: raw
14 material for evolution. *Molecular Ecology* **15**, 1197-1211.

15 Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene
16 regulation. *Nature* **430**, 85-88.

17 Wittkopp PJ, Haerum BK, Clark AG (2006) Parent-of-origin effects on mRNA expression in
18 *Drosophila melanogaster* not caused by genomic imprinting. *Genetics* **173**, 1817-1821.

19 Wittkopp PJ, Haerum BK, Clark AG (2008) Independent effects of *cis*- and *trans*-regulatory
20 variation on gene expression in *Drosophila melanogaster*. *Genetics* **178**, 1831-1835.

21 Wittkopp PJ, Haerum BK, Clark AG (2008) Regulatory changes underlying expression
22 differences within and between *Drosophila* species. *Nature Genetics* **40**, 346-350.

23 Wray GA (2007) The evolutionary significance of *cis*-regulatory mutations. *Nature Reviews*

1 *Genetics* **8**, 206-216.

2

1

2 **Acknowledgments:**

3 We thank the FAS Center for Systems Biology for the use of their Facility and the
4 Sequencing Platform at the Broad Institute. PF was supported by a Marie Curie Outgoing
5 Fellowship from the European Commission and postdoctoral fellowships from the Swiss
6 National Science Foundation. CRL was supported by the National Sciences and
7 Engineering Research Council of Canada (NSERC) and a Frank Knox Memorial
8 Fellowship at Harvard University, and a Canadian Institute of Health Research (CIHR)
9 Fellowship at University of Montreal and is now a CIHR New Investigator at Laval
10 University. Funding for this work was provided NIH grants GM068465 and GM065169
11 to DLH and a March of Dimes Basil O'Connor Starter Scholar award to PJW (5-FY07-
12 181). P.J.W. is an Alfred P. Sloan Research Fellow.

1 **Figure captions:**

2

3 **Figure 1: High-throughput sequencing technology allows measurement of relative**
4 **allelic expression genome-wide.**

5 The schematic representation shown illustrates the steps require to collect allelic
6 expression data. Key parameters associated with each step that ultimately affect the
7 statistical power for detecting significant AI are also shown.

8

9 **Figure 2: The expected proportion of informative reads increases with genetic**
10 **divergence and read length.**

11 (A, B) Black lines show expected proportions of informative reads (i.e. sequence
12 fragments that could be unambiguously assigned to one allele) predicted by Eq. 3 for
13 transcribed sequences containing 0.1, 0.5, 1, or 5% sequence divergence, as indicated.
14 Predictions are shown in which either one SNP (A) or two SNPs (B) were required for a
15 sequencing read to be informative for measuring allelic expression. (C, D) Predictions
16 based on 0.1% and 1% sequence divergence and requiring only one SNP to be
17 informative are shown again, as they were in (A). Results from simulated datasets are
18 also shown. Each simulation contained either 20 (C) or 200 (D) reads that were generated
19 using a virtual 2000 bp mRNA sequence, 0.1% or 1% sequence divergence, and
20 sequencing reads of 35, 150, 300, and 800 bp. Each scenario was simulated 500 times,
21 and is summarized by boxplots showing the median, lower and upper quartiles, as well as
22 the 1.5 interquartile range. The gray lines are the 95% confidence intervals of the

expected proportions based on binomial sampling (Clopper-Pearson interval on Eq. 1, Clopper & Pearson 1934).

Figure 3: Predicted proportions of genes with more than 200 informative reads for a given sequencing depth are consistent with simulated data.

Predicted values (lines) were obtained using Eq. 14, assuming a mean read length of 150 bp and sequence divergence of 0.1%, .5%, 1%, and 5%, as indicated. Simulated data (points) used distributions of transcript abundance, read length, and sequence divergence, as shown for the insets within Fig. 3. Two replicate simulations were performed and found to be highly correlated with each other (Spearman's Rho > 99%).

Figure 4: Detecting significant allelic imbalance for genes with small differences in allelic expression requires a large number of informative reads per gene.

Statistical power for detecting significant allelic imbalance for a Type I error $\alpha = 5\%$ under different conditions is shown. Each line shows the power to detect significant AI, assuming that the true value of AI is 1, 1.25, 1.5, 2, 7, 10, or 100..

Figure 5: Read sampling strategy affects the proportion of informative reads per gene and thus the number of genes for which significant AI can be detected.

(A, B) Simulated proportions of genes with more than 200 informative reads using a random or a targeted read sampling strategy are shown for mean read lengths of 35 bp (A) and 150 bp (B), with individual reads sampled from a Poisson distribution. See Fig. 2 for a more detailed description of the simulation parameters. (C) The proportion of

1 informative reads per gene using random (left), targeted (middle) , and mixed (right)
2 sampling strategies are shown. Each beanplot represents the distribution (500 replicates)
3 of the proportion of informative reads among 500 sampled reads . The horizontal bar
4 shown on each beanplot indicates the mean of these distributions. For the mixed strategy,
5 fragments with sequence lengths drawn from a Poisson distribution with a mean of 500
6 bp were anchored to a fixed, predetermined location (the 3' end), and sequences of either
7 18 or 75 bp were taken from each end to simulate paired-end sequencing.

8

9 **Figure 6: Analysis of allelic expression and AI in *Drosophila* F₁ hybrids.**

10 (A) The distribution of sequence fragment lengths for both informative and uninformative
11 reads is shown. (B) The number of genes in different gene expression level classes (as
12 measured by the abundance of informative reads) are shown along with the number of
13 genes in each class that showed significant AI. (C) The top panel shows the proportion of
14 genes with significant AI (see Supplementary Table S1) for which the *D. melanogaster*
15 allele is most abundant. The bottom panel shows the proportion of informative reads in a
16 given expression level class that were assigned to *D. melanogaster*. In both panels, the
17 dotted line corresponds to a balanced proportion (50%). (D) The relationship between
18 relative allelic expression as measured by 454 sequencing and by pyrosequencing is
19 shown. For pyrosequencing, the average of eight replicates is plotted and the 95%
20 confidence intervals are indicated by the horizontal bars. For 454 sequencing, the relative
21 number of informative reads is shown, with vertical bars indicating the Clopper-Pearson
22 95% confidence intervals derived from binomial sampling (see Supplementary Fig. S3).
23 The dotted line indicates the slope of the non-parametric regression.

1

2

3 **Figure S1: Alternative model for predicting the proportion of informative reads:**
 4 **incorporating the finite length of mRNA sequence .**

5 Equations described in the main text treat mRNA length as infinite despite the fact that
 6 real mRNA molecules have a finite length. To examine the impact of this assumption on
 7 the accuracy of predictions, we replaced the binomial distribution used for sampling
 8 SNPs within an mRNA with a hypergeometric distribution that allows the length of an
 9 mRNA to be considered explicitly. Black solid lines show expected proportions of

10 informative reads predicted by Eq. 3 with $\Pr(X \geq x) = \sum_{k=x}^l \frac{\binom{l_g d}{k} \binom{l_g (1-d)}{l-k}}{\binom{l_g}{l}}$, where l_g is the

11 mRNA length. The black dashed lines are expected proportions of informative reads
 12 predicted by Eq. 3 using Eq. 1 (assuming an infinite mRNA length). The gray lines are
 13 the 95% confidence intervals of the expected proportions. $l_g = 2000$ bp (see Fig. 2 for
 14 details).

15

16 **Figure S2: The location of a SNP within a sequence affects its probability of being**
 17 **sampled in a random sampling strategy.**

18 The sampling probability of each nucleotide position in a 2000 bp mRNA sequence is
 19 shown. (A) The figure shows edge effects of read sampling: bases close to the 5' or 3'
 20 ends of transcripts had a lower probability of being observed among 10,000 reads, each
 21 50 bp length. The red line shows the average probability of sampling a SNP at any site.

1 The green line shows the expected maximum probability (see text). Panels (B), (C), and
2 (D) show variable sampling schemes.

3

4 **Figure S3: The number of informative reads per gene, as well as the magnitude of**
5 **AI, affect the accuracy of AI estimates.**

6

7 **Figure S4: Distribution of significant AI in F₁ hybrids.**

8 Genes showing greater abundance of *D. melanogaster* and *D. simulans* alleles are
9 positive and negative, respectively. The X-axis shows the AI values after a log (base 2)
10 transformation.

11